JOURNAL OF APPLIED BEHAVIOR ANALYSIS 2011, **44,** 571–586 NUMBER 3 (FALL 2011)

CALIBRATION OF OBSERVATIONAL MEASUREMENT OF RATE OF RESPONDING

OLIVER C. MUDFORD

UNIVERSITY OF AUCKLAND, NEW ZEALAND

AND

JASON R. ZELENY, WAYNE W. FISHER, MOLLY E. KLUM, AND TODD M. OWEN

MUNROE-MEYER INSTITUTE AND

UNIVERSITY OF NEBRASKA MEDICAL CENTER

The quality of measurement systems used in almost all natural sciences other than behavior analysis is usually evaluated through calibration study rather than relying on interobserver agreement. We demonstrated some of the basic features of calibration using observer-measured rates of free-operant responding from 10 scripted 10-min calibration samples on video. Five novice and 5 experienced observers recorded (on laptop computers) response samples with a priori determined response rates ranging from 0 to 8 responses per minute. Observer records were then compared with these predetermined reference values using linear regression and related graphical depiction. Results indicated that all of the observers recorded rates that were accurate to within ±0.4 responses per minute and 5 were accurate to within ±0.1 responses per minute, indicating that continuous recording of responding on computers can be highly accurate and precise. Additional research is recommended to investigate conditions that affect the quality of direct observational measurement of behavior.

Key words: calibration, continuous recording, observer accuracy, observer precision, recording and measurement

Continuous measurement of free-operant responding by observers recording on handheld or laptop computers has become more common in applied behavior analysis research than discontinuous recording (e.g., momentary time sampling or partial-interval recording; Mudford, Taylor, & Martin, 2009). A benefit for applied behavior analysis, as a natural science of human behavior, is that the products of measurement can be reported in conventional units (e.g., responses per minute) from recording of behavioral events.

Behavior analysts who present their data in conventional units of measurement have the opportunity to bring their methods for assessing and reporting the quality of those data in line with other natural sciences. We use the overarching term *quality* to include concepts of accuracy, precision, and errors of measurement from the science of measurement (metrology) (e.g., Hauck, Koch, Abernethy, & Williams, 2008; The Royal Society of Chemistry, 2003); interobserver agreement from applied behavior analysis (e.g., Mudford, Taylor, et al., 2009); and reliability and validity from social sciences (e.g., Gresham, 2003). Metrological concepts and interobserver agreement will be defined and described.

According to Russell (1937), ''measurement demands some one-one relations between the numbers and magnitudes in question'' (p. 176).

Oliver Mudford was on research leave at the Munroe-Meyer Institute at the time of the study. We are grateful to the observers who volunteered to contribute to the research and to Kasey Stephenson and Kelly Bouxsein for scheduling observations. The first author thanks Ivan L. Beale (formerly of the University of Auckland) for encouraging his interest in calibration that started in 1987.

Correspondence concerning this article should be addressed to Oliver Mudford, Applied Behaviour Analysis Programme, Department of Psychology (Tamaki Campus), University of Auckland, PB 92-109, Auckland 1142, New Zealand (e-mail: o.mudford@auckland.ac.nz).

doi: 10.1901/jaba.2011.44-571

The accuracy of measurement is the extent to which differences in the numbers assigned to various magnitudes reflect, or relate to, the actual differences in those magnitudes. For a greengrocer, accurately measuring the weight of fruits sold is important to making a living and maintaining good customer relations. For a chemist, accurately measuring elements of a compound is essential to maximize the beneficial effects and to minimize the untoward effects of that compound. For an applied behavior analyst, accurately measuring the target response is critical to data-based decision making: Are the data accurate enough to interpret responsibly the obtained results from assessment and intervention?

In other natural sciences, the accuracy of a measurement instrument is typically determined by comparing or calibrating the instrument's measurements with known standards (e.g., testing the accuracy of a balance scale by measuring a set of objects of known weight). Although accuracy has long been recognized as the gold standard for assessing the quality of observational measurement (e.g., Cooper, Heron, & Heward, 1987, 2007; Johnston & Pennypacker, 1980, 1993, 2009), applied behavior analysts have rarely assessed the accuracy of their continuous behavioral measurement systems (Mudford, Martin, Hui, & Taylor, 2009; Mudford, Taylor, et al., 2009).

As a substitute for accuracy, interobserver agreement has been reported. Interobserver agreement is computed by comparing two continuous records that independent observers recorded contemporaneously. Interobserver agreement may be considered a poor surrogate for accuracy because we cannot determine the extent to which either of the observer's records represents a ''true'' account of the behavior of interest. Nevertheless, the use of interobserver agreement computation methods is considered to be indispensible for ensuring the specificity of behavioral definitions as they are refined during the initial development of an observational system, ensuring that observers are responding homogeneously to defined behavioral responses, and assessing the effects of observer training.

A recent study demonstrated the use of three common interobserver agreement algorithms to assess accuracy by comparing observers' continuous records of responding with criterion records of the observed samples (Mudford, Martin, et al., 2009). The outcome was unsatisfactory because the algorithms all showed systematic bias. Block-by-block and exact agreement methods tended to inflate accuracy at lower rates of responding, exact agreement reduced apparent accuracy at higher rates, and time-window analysis inflated accuracy estimates at higher rates. In other words, none of these algorithms were found to be inarguably preferred.

The overwhelming problem with attempting to assess observer accuracy with interobserver agreement algorithms can be deduced from considering the International Standards Organisation's International Vocabulary of Metrology definition of *accuracy* as ''Closeness of agreement between a measured quantity value and a true quantity value of a measurand'' (quoted in Hauck et al., 2008, p. 841). The measurand, or that which is measured, for free-operant responding in applied behavior analysis is usually a simple summary statistic from an observation session (e.g., responses per minute or percentage duration; for some exceptions, see Fahmie & Hanley, 2008). Interobserver agreement algorithms do not address directly the typical measurand. However, they are helpful for assessing within-session accuracy (i.e., the *process* of measurement), even if not the accuracy of the substantive data (i.e., the *product* of measurement).

Johnston and Pennypacker (1980, 1993, 2009) and Cooper et al. (1987, 2007) have consistently recommended that applied behavior analysts apply calibration methods to address the issues concerning the accuracy of

behavioral data. Johnston and Pennypacker (2009) defined *calibration* as "Evaluating the accuracy and reliability of data produced by a measurement procedure and, if necessary, using these findings to improve the procedure" (p. 148). The recommendation to adopt calibration has not been taken up by researchers in applied behavior analysis.

The process of calibration in natural sciences involves comparing measurements made by the instrument to be calibrated with accurate values of the measurand. It should be highlighted that a pragmatic definition of accuracy can be adopted. For example, applied chemists agree that insistence on including the philosophically problematic concept of "true" values of measurands in the definition of accuracy is no longer essential. For example, the Royal Society of Chemistry definition states that *accuracy* is "The closeness of agreement between a test result and the accepted reference value" (The Royal Society of Chemistry, 2003, p. 1; see also Hauck et al., 2008). In regard to engineering instrument calibration,

> The term *true value*, then, refers to a value that would be obtained if the quantity under consideration were measured by an *exemplar method*, that is, a method agreed upon by experts as being sufficiently accurate for the purposes to which the data will ultimately be put. (Doebelin, 1966, pp. 41–42)

Comparison between obtained and reference values is usually quantified in natural sciences using simple linear regression analysis. Many software packages (e.g., Excel, Sigmaplot) can be used for the statistical analysis and its graphical representation. Linear regression can best be visualized graphically by first plotting the independent variable (reference values) from the *x* axis and the dependent variable (obtained value) from the *y* axis. Second, a regression line is plotted that provides a line of best fit through the data points on the graph by the least squares method. Third, provided the assumptions behind the use of linear regression are not violated by the data, the *accuracy* of the instrument can be judged by the closeness of the slope of the line to the diagonal. If, for example, the regression line overlays the diagonal, the instrument is perfectly accurate. Fourth, the *precision* of the instrument can be assessed by the closeness of all points to the regression line. If all points lie on the regression line, the instrument is perfectly precise. Accuracy and precision are independent and by no means synonymous: An instrument can be accurate but imprecise, and vice versa. Readers are referred to textbooks concerning natural science analytical methods for details of calibration (e.g., Burke, 2001; Hibbert & Gooding, 2005; The Royal Society of Chemistry, 2006).

Calibration involving regression models is ubiquitous in science. The following are several disparate examples from among hundreds located in a search of scientific research articles. Examples from applied physics include (a) calibration of equipment to determine the quality of a variety of foods, including olive oil, milk, salami, and chocolate (Niemoller & Behmer, 2008); and (b) calibration of equipment to measure the age, rate of expansion, and properties of our universe (Feigelson & Babu, 1992; Grimm, Gilfanov, & Sunyaez, 2003). Applied biology uses regression models for calibration (e.g., calibration of medical measurement instruments; Yang et al., 2008), and chemists calibrate their laboratories and analytical measuring equipment and processes by comparing obtained data from their analysis of chemical solutions with known concentrations using regression analysis (e.g., Burke, 2001; Garofolo, 2004; Hibbert & Gooding, 2005; Kramer, 1998). The known concentrations are the "accepted reference values" referred to in modern definitions of accuracy.

Although we did not locate any applied behavior-analytic research articles on the topic of calibration, direct observations of behavior have been calibrated in studies of animal behaviors (e.g., Holland, Dabelsteen, Bjorn, & Pedersen, 2001; Lee & Hockey, 2001). An

example of the use of regression models for calibration of measurements made by sciences that do (or might) have direct relevance to applied behavior analysis concerned apparatus for measuring human eye-tracking movements (Hong, Avila, Wonodi, McMahon, & Thaker, 2005).

In sum, (a) calibration is used across the natural sciences; (b) regression analyses are employed in the process of calibration; and (c) despite general distaste for correlational methods, behavior analysts need not fear being branded unscientific by considering regression models for assessing the quality of their data.

For applied behavior analysis, exemplar methods for obtaining true reference values against which observers' recordings could be calibrated derive from criterion records. Methods for creating criterion records have included electromechanical recording (e.g., Kapust & Nelson, 1984), using predetermined scripted performances with scripts acting as true records (e.g., Lerman et al., 2010; Powell, Martindale, Kulp, Martindale, & Bauman, 1977), repeated viewing of video records by more than one observer until consensual agreement has been achieved (e.g., Boykin & Nelson, 1981; Mudford, Martin, et al., 2009), and relying on an expert observer who has been trained to high interobserver agreement levels (e.g., Sanson-Fisher, Poole, & Dunn, 1980; Wolfe, Cone, & Wolfe, 1986).

The primary purpose of the current study was to illustrate the potential benefits of calibration using simple linear regression with response rates obtained from observers' records of free-operant behavior. A secondary purpose was to assess differences between experienced and novice observers in terms of their accuracy, error, and precision. The measurement system assessed was at the level of individual observers using the same equipment, data-recording software, observational protocols, and calibration samples.

METHOD

Participants

Two groups of observers, experienced and novice, were recruited from the staff of a university-based treatment and research center specializing in developmental disorders. The experienced group consisted of three male and two female adult staff members who work in a problem behavior assessment and treatment unit. Four had bachelor's degrees and one had a master's degree. Experienced observers had been employed at the center for an average of 31 months (range, 6 to 48 months), and collected data daily with the data-collection system used in this study. Their initial training had been through lecture, observation of others' data collection with commentary, guided practice, and recording with feedback. They were considered to be trained following three consecutive in vivo sessions from which interobserver agreement exceeded 80% for all target behaviors and salient environmental events.

The novice group of observers, four women and one man, had been employed for 2 days to 28 months (M = 8 months) in an early intervention unit. They had not previously used the behavioral measurement system described in this report. Two novices had bachelor's degrees, and three were undergraduates.

Responses Recorded and Definitions

Observers recorded the occurrence of three behaviors: c*lient pinches therapist,* defined as the client's hand moving towards the therapist's upper arm to within 2.5 cm; *model prompt*, defined as the therapist placing one block on top of another; *verbal and gestural prompt*, defined as the therapist saying ''stack'' and pointing to blocks simultaneously.

Observational Materials (Calibration Samples)

As a rule of thumb in analytical chemistry, six samples with reference values distributed across the full range of intended measurements are considered to be the minimum for

calibration purposes, although 10 are usually preferred (Hibbert & Gooding, 2005; Kramer, 1998). Garofolo (2004) recommended six to eight nonzero samples plus a zero (or "blank") sample. The range of rates of responding in the samples we used (0 to 8 responses per minute) was chosen because it encompasses approximately 90% of all values for responses per minute reported in the *Journal of Applied Behavior Analysis* from 1998 through 2007 from continuous measurement of free-operant responding (Mudford, Locke, & Jeffrey, 2011).

We produced 10 video samples (10 min each) that contained role-played behaviors. Samples were recorded with a JVC MiniDV digital video camera. Two adults playing the roles of a therapist and a client sat on adjacent sides of a table (130 cm by 130 cm) in a clinic room. The tabletop contained approximately 25 Lego blocks, which both role players manipulated. Video samples started with the instruction "one, two, three, go," at which point the role players started timers. Scripts were taped to the tabletop with prearranged times (in minutes and seconds) beginning at the start of a session. Role players were scheduled to display the responses to be measured at the prearranged times. The video recordings were edited to have 5 s before the session started on "go," and the end of the sample, 600 s later, was signaled by a screen announcing, "end of session."

The scripts used in the role plays varied in the number of pinches that occurred: 0, 5, 10, 20, 30, 40, 50, 60, 70, and 80. The distribution of pinches across 10-min sessions was derived from real client problem behavior obtained from a search of deidentified clinical archives. The therapist's script contained 10 modeled prompts and five verbal or gestural prompts in every sample. Distribution of therapist responses across the 600-s samples was determined separately for each sample from a random number generator (www.random.org).

Criterion Records

The time of occurrence of all recorded responses, to 1-s precision, was obtained from scrutiny of the completed video samples. The first, fourth, and fifth authors independently noted the times of occurrences through playing, pausing, rewinding, and replaying the video recordings using Microsoft Windows Movie Maker software (Version 6.0, 2007). The only discrepancies between independent records were on the exact second in which responses occurred. Discrepancies were resolved by authors reviewing the sample together until agreement on time of occurrence was achieved. For 31 of 515 occurrences (6%) of target behaviors, the criterion records were finalized after adjustment by ± 1 s. The agreed records were converted to computer files compatible with the data-analysis software employed.

Observational Setting, Equipment, and Recording Software

Observers, one at a time, recorded responses from five video samples in a session, with a 5-min break provided between samples. Longer breaks ranging from 45 min to 2 days occurred between sessions, depending on the participants' work schedules. Recording took place in a quiet room (3 m by 4 m). A Dell Latitude D630 laptop was connected to a 38-cm Entuitive video monitor positioned above and behind the laptop. The observer sat 130 to 150 cm from the video monitor. Video samples were stored on separate HP 2GB USB 2.0 flash drives labeled with a letter denoting the sample. Video was played from the flash drive onto the upper monitor. The laptop monitor showed the recording screen for DataPal data-recording software (available from UNeMed@unmc.edu). Observers pressed the laptop's keyboard keys to record the start of an observation (tab key), pinches (A), model prompts (B), verbal or gestural prompts (V), and end of session (Ctrl-E).

Procedure

Before the observers' first recording, they were shown an introductory video with examples of the behaviors they would be recording, their definitions, and recording instructions. Also, they viewed a 2.5-min video showing the setting for the videos (i.e., client and therapist at the table). The observers had been informed that the study concerned the accuracy of observational measurement. Observers' records of model and verbal prompts were not analyzed; they were distracters. Observers were not informed that their accuracy of recording codes B and V would not be examined.

Each observer recorded video samples in a unique random sequence. Observers were provided with one video sample (i.e., one flash drive) at a time to plug into a USB port on the laptop. Instructions to observers were to start the session when they heard "go," record the target responses as they occurred, and end the session when instructed by the video monitor. After each sample, they exchanged the flash drive for the next one in their random sequence.

Analyses

Accuracy assessment using interobserver agreement algorithms. Agreement between observer's records and criterion records was computed using block-by-block agreement, exact agreement, and time-window analysis algorithms (Mudford, Taylor, et al., 2009). The options selected in the Obswin 32 analysis software (www.antam.co.uk/obswin.htm) were 10-s intervals imposed on the data streams for block-by-block and exact agreement, and ± 2 s tolerance for agreement permitted time-window analysis.

Calibration analysis. We performed calibration analysis on rates of responding derived from observers' records of pinching (Code A) only from the 10 samples. Counts of the occurrence of Code A from observers' whole-session summaries divided by the recorded session durations (in minutes) gave rates measured by each observer. Criterion record rates (i.e., reference values) were computed in the same manner. The criterion record rate was the independent variable (x-axis values), and observers' records were the dependent variable (y-axis values) in linear regression using the least squares method computed with Sigmaplot Version 11 (Systat Software Inc.). Prediction intervals at the 95% confidence level for prediction of Y from X were computed and graphed alongside the regression lines to enable visual analysis of individual observer's accuracy and precision.

RESULTS

Accuracy Assessment Using Interobserver Agreement Algorithms

Table 1 shows the mean and range of accuracy computed by percentage agreement for each observer with typically used algorithms (block-by-block agreement, exact agreement, and time-window analysis). The means for experienced observers exceeded 96% agreement on all measures. No experienced observer's recording on any individual session was less than 90% accurate by any interobserver agreement measure. Novice observers showed more variation in accuracy (ranges, 89% to 96.9% for means, 56.8% to 100% for individual sessions). The novices' means overlapped with the experienced observers somewhat in block-by-block agreement, less so in exact agreement, and the least with time-window analysis. This finding indicates that novices, in general, were somewhat less accurate in responding to occurrences of pinching or with the timing of their recording to ± 2 s of events in the criterion records (i.e., they were more delayed in pressing the A key following pinches).

Calibration Analysis

Statistical data. Table 2 presents statistics from the linear regression analysis. The linear calibration model is expressed by the formula $Y = a + bx$, where Y is the product of the observer's measurement (in responses per min-

Table 1
Accuracy by Common Percentage Agreement Algorithms for all Observers

Observers		Block-by-block agreement mean (range)	Exact agreement mean (range)	Time-window analysis ±2 s mean (range)
Experts	WJ	98.1 (95.6 to 100)	97.5 (93.3 to 100)	96.4 (90.0 to 100)
	JC	99.4 (97.5 to 100)	99.2 (96.7 to 100)	99.5 (95.2 to 100)
	FL	98.9 (96.1 to 100)	98.3 (95.0 to 100)	98.9 (93.3 to 100)
	DB	99.4 (97.5 to 100)	98.8 (96.7 to 100)	99.8 (98.0 to 100)
	DH	98.8 (93.3 to 100)	98.2 (91.7 to 100)	98.8 (90.4 to 100)
Novices	DK	97.9 (93.3 to 100)	96.7 (93.3 to 100)	95.2 (66.7 to 100)
	KA	98.2 (94.2 to 100)	97.2 (93.3 to 100)	96.9 (86.7 to 100)
	MW	98.5 (92.2 to 100)	97.7 (86.7 to 100)	94.7 (66.7 to 100)
	DJ	97.4 (92.5 to 100)	96.3 (91.7 to 100)	92.7 (58.3 to 100)
	PM	95.1 (87.8 to 100)	92.7 (80.0 to 100)	89.0 (56.8 to 100)

ute) from a sample (dependent variable); x is the rate (in responses per minute) from the criterion record (independent variable); a is the value (in responses per minute) of the intercept on the y axis (i.e., the value of Y when x is zero); and b is the slope of the regression line, otherwise known as measurement sensitivity. Ideally, the intercept value (a) should be close to zero, which it was for all observers. Observer FL's intercept value for a was most discrepant from zero, but still not significantly different from zero ($p = .43$). The values for slope (b) should be close to 1.0 for a good measurement system, with accuracy (i.e., absence of systematic error; Hauck et al., 2008; The Royal Society of Chemistry, 2003) declining with increasing divergence from 1.0. With the present data set, slopes were all close to but slightly less than 1.0, indicating a tendency for observers to underrecord the occurrence of events or overrecord the observational session duration by terminating their record with Ctrl-E with less than exact timing. Six observers, of whom four were experienced, were close to perfectly accurate, with $b \geq 0.996$ (Table 2, column labeled "slope").

R^2, although typically reported with linear regression, is said not to be a useful statistic when associated with a regression line in calibration because high values are to be

Table 2
Calibration Statistics for All Observers

Observers		Calibration statistics				
		a Intercept	b Slope	R^2	Standard error of estimate	Predicted rate at criterion rate of 6.5 ± 95% CI
Experts	WJ	−0.0273	0.986	.999	0.081	6.4 ± 0.2
	JC	0.0182	0.996	1.000	0.033	6.5 ± 0.1
	FL	−0.0336	0.996	.999	0.075	6.5 ± 0.1
	DB	−0.0039	0.997	1.000	0.035	6.5 ± 0.1
	DH[a]	−0.0016	0.996	1.000	0.032	6.5 ± 0.1
Novices	DK	0.0054	0.996	1.000	0.050	6.5 ± 0.1
	KA[b]	−0.0247	0.996	.997	0.158	6.4 ± 0.4
	MW	−0.0193	0.974	.999	0.099	6.4 ± 0.2
	DJ[c]	0.0027	0.974	.999	0.106	6.3 ± 0.3
	PM	0.0046	0.982	.998	0.119	6.4 ± 0.3

[a] Data shown in Figure 1.
[b] Data shown in Figure 2.
[c] Data shown in Figure 3.

expected, with an R^2 of .999 often being cited as the minimum acceptable in analytic chemistry (Hibbert & Gooding, 2005). Instead, the standard error of the estimate (*SE*) should be reported to describe the variability about the regression line. The standard error of the estimate can be described as a sensitive measure of precision. The further the value is from zero, the greater the random error in the measurement system; therefore less precision. Table 2 shows that four of the five lowest standard errors (i.e., highest precision observers) were obtained from the experienced observers.

Visual presentation of regression statistics. Although researchers and practitioners in many natural sciences are trained to conduct calibration analyses, and many may be able to obtain all the information they need from statistics (e.g., Table 2), behavior analysts in general, along with many in other natural sciences, find graphical presentation helpful when interpreting regression analyses. To illustrate, we show regression lines for three of the participants: (a) Participant DH (Figure 1), who was highly accurate (slope = 0.996) and the most precise observer (*SE* = 0.032); (b) Participant KA (Figure 2), who was the least precise observer (*SE* = 0.158); and (c) Participant DJ (Figure 3), who had the lowest accuracy (slope = 0.974; tied with Participant MW) and also low precision (*SE* = 0.106). At the left extreme of the regression lines in the top panel of each figure, all lines pass very close to the origin ($x = 0$, $y = 0$) as would be expected from small values of *a*, not significantly different from zero. At the right end of the regression lines for DH and KA, the line ends at close to $x = 8$, $y = 8$, reflecting values for *b* (slope) close to 1.0. For DJ, however, the slope (0.974) is flatter, and it can be seen that when $x = 8$, $y < 8$.

Up to this point many readers will have found descriptions and illustrations of linear regression redundant with their background in psychological research methods from their undergraduate or graduate studies. What follows may be unfamiliar to many *JABA* readers because calibration has not been used in applied behavior analysis and, to our knowledge, is not often taught to psychologists in graduate schools.

In the bottom panel of each figure, the dotted lines above and below the regression lines are plots of the 95% confidence intervals for predictions about new observations from data derived from observations conducted for calibration. Confidence bands for prediction are not the same as the more familiar confidence intervals about the regression line that describe the extent of variability of data. Although the standard error of the estimate is used in the computation of both types of confidence intervals, the values for prediction confidence intervals are always farther from the regression line. The reason why they are shown in Figures 1, 2, and 3 is that calibration is used for prediction from (not merely description of) the parameters of regression. It can be seen that predictions from DH's regression (Figure 1) can be more precise, because the confidence intervals are narrower, than from DJ's regression (Figure 3). KA was least precise, with the widest of the three participants' confidence bands for prediction (Figure 2). Comparisons among other observers can be made by examining standard error values in Table 2.

The purpose of calibration illustrated. From the information presented thus far, it is possible to predict how observers would perform (with 95% confidence) when measuring pinching behavior at rates that are within the bounds of the calibration samples (i.e., from 0 to 8 responses per minute). The lower panels in Figures 1, 2, and 3 present detail showing the prediction of observers' performance when measuring a criterion rate of 6.5 responses per minute. This criterion rate was selected somewhat arbitrarily. Confidence intervals are not exactly parallel with the regression line, being narrowest at the mean of *x* values (3.7) and widest at the extremes (e.g., 7.9). Nevertheless,

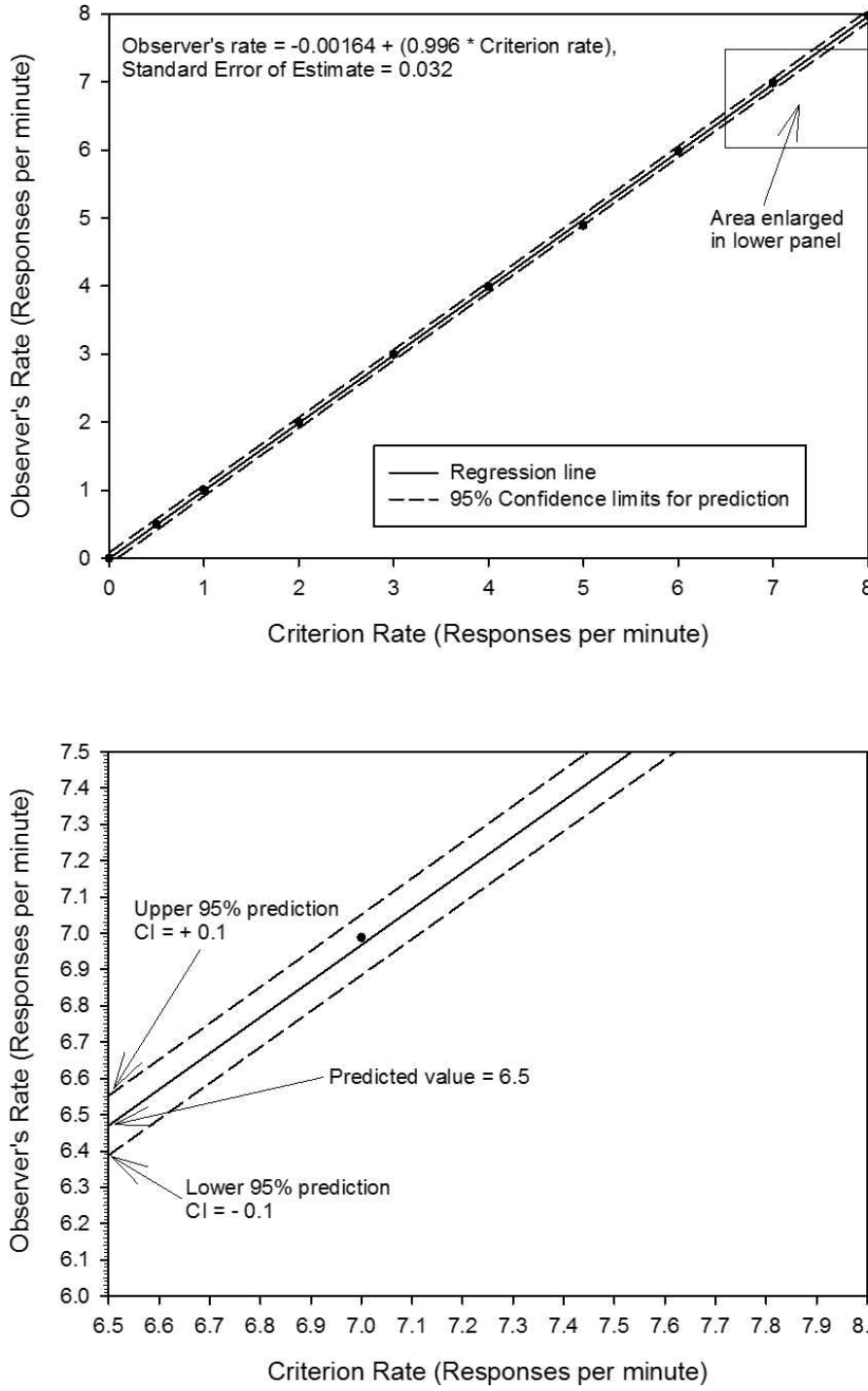

Figure 1. An accurate precise observer, DH. Regression line and confidence intervals (CI) for prediction. Lower graph is enlarged detail from upper graph showing predicted observer's rate at criterion rate of 6.5 responses per minute.

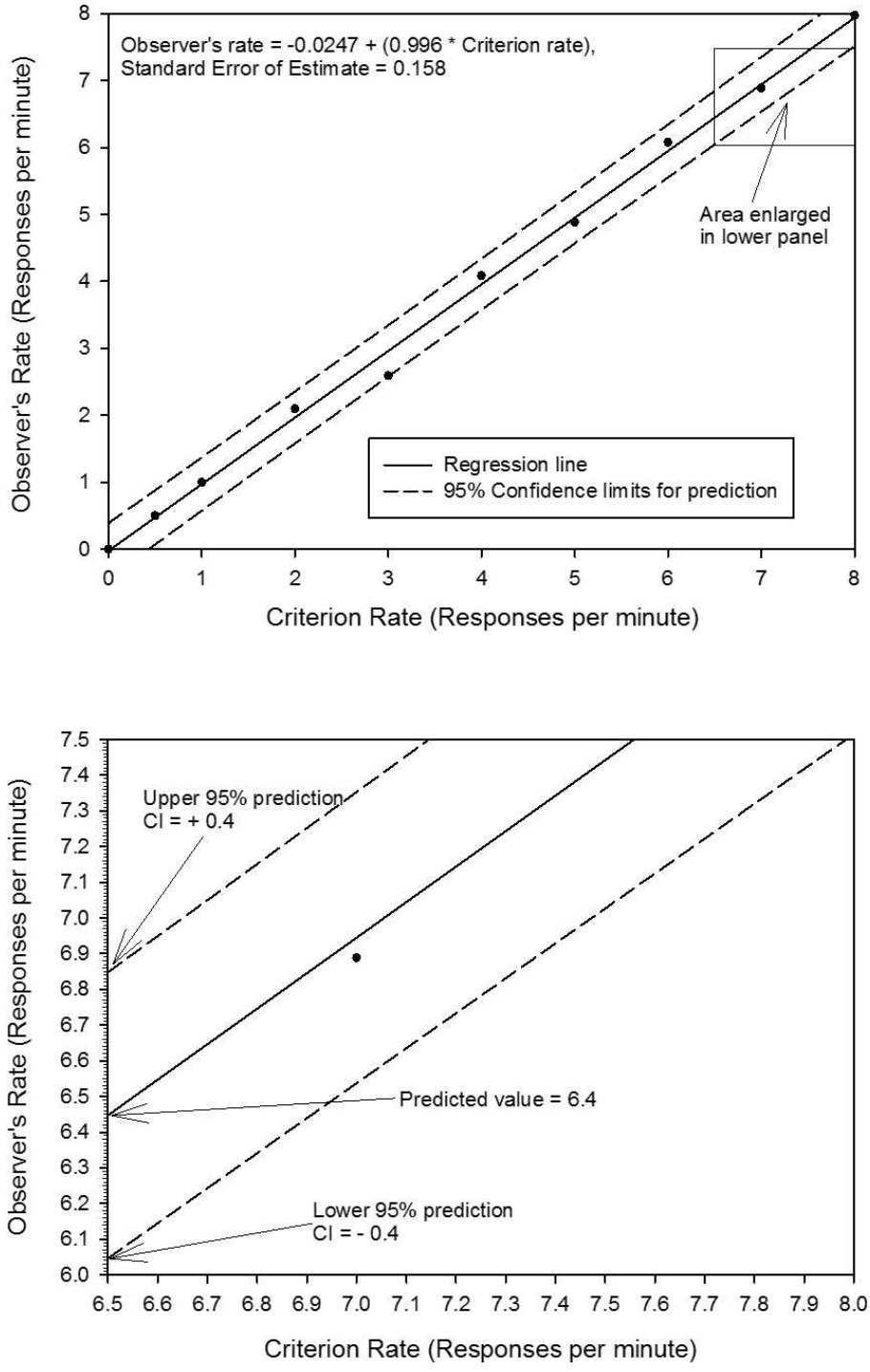

Figure 2. The least precise observer, KA. Regression line and confidence intervals (CI) for prediction. Lower graph is enlarged detail from upper graph showing predicted observer's rate at criterion rate of 6.5 responses per minute.

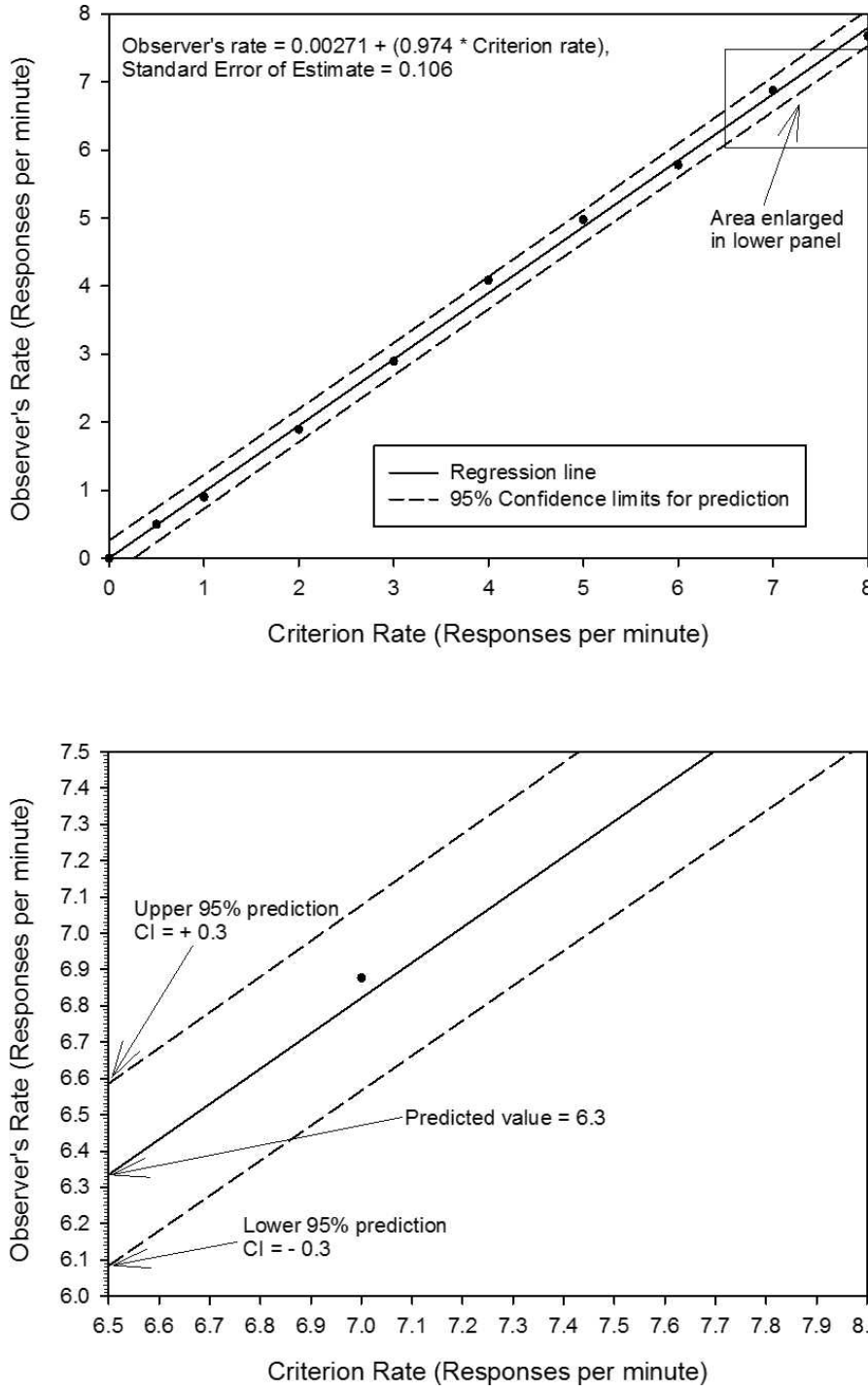

Figure 3. The least accurate observer, DJ. Regression line and confidence intervals (CI) for prediction. Lower graph is enlarged detail from upper graph showing predicted observer's rate at criterion rate of 6.5 responses per minute.

limits of prediction for experienced observers did not change from their values (rounded to one decimal place) from 6.5 (Table 2) to 7.9 responses per minute. The predicted value can be read from the *y* axis at the point of intersection of the regression line when $x = 6.5$. The precision of that value can be read from the *y* axis by measuring the width of the band bounded by the confidence limits (in responses per minute), divided by 2, and expressed as ± responses per minute. DH was the most accurate of the three observers shown, with a predicted value of 6.5 responses per minute and with more precision (± 0.1 responses per minute). KA was less accurate at 6.4 responses per minute and was the least precise (± 0.4 responses per minute). The graph for DJ shows the least accurate prediction of 6.3 ± 0.3 responses per minute.

It should be noted that levels of error around the prediction can be computed statistically. However, that it is acceptable to use graphical means in other natural sciences (e.g., analytical chemistry; Burke, 2001) may encourage behavior analysts to adopt the method we have used in this illustration.

DISCUSSION

We have shown, statistically and graphically, how the methods and interpretation of calibration can be applied to assess the quality of direct observational measurement of response rates. The two sources of error of greatest interest in calibration are systematic errors and random errors. *Accuracy* is defined as the absence of systematic error (neither consistently overestimating nor underestimating the true reference values; The Royal Society of Chemistry, 2003). *Precision* is defined as the absence (or minimization) of random error in measurement (e.g., Bragg, 1974). Experienced observers demonstrated almost uniformly high levels of agreement with criterion records, accuracy, and precision. We can conclude from calibration analysis that four of five experienced observers, using DataPal to record from the calibration samples studied, can measure rates of responding up to 8 responses per minute with nearly perfect accuracy (to one decimal place) and precision within ±0.1 responses per minute. The fifth experienced observer (WJ) underrecorded events, but a predicted difference of 0.1 responses per minute when the criterion rate was 6.5 responses per minute is a level of error equaling just 1.5% of the criterion value (i.e., $0.1 \div 6.5 = 1.5\%$), which may be considered sufficiently accurate for most interpretations of behavioral data.

Novice observers, as a group, did not fall far behind the experienced observers in accuracy and precision. One novice (DK) was found to be as accurate and precise as the best experienced observers. With training similar to that described for the experienced observers, it may be reasonable to expect that novices would all become experts. More targeted training may be suggested from calibration results (e.g., systematic errors leading to underrecording suggest that closer attention to the recording task might be encouraged). Note that we recommend training observers in the first instance, not applying post hoc correction of systematic biases to the obtained values based on calibration of individual observers.

We must emphasize that the results of this calibration study cannot be generalized or extrapolated beyond the particular measurement systems assessed, which involved experienced and novice observers watching videos and recording the data on laptop computers using the DataPal software. Thus, results are applicable only to the individual participants, in the setting described, and with the equipment, response definitions, procedures, and calibration samples used. Consequently, our data should not be interpreted to claim that, for instance, (a) continuous recording of responding is as accurate and precise with other definitions, equipment, procedures, and so on; (b) observers in the present study would be

highly accurate with other definitions, and so on; or (c) given identical methods, observers would produce the same levels of accuracy at rates outside the range of calibration samples (i.e., $>$ 8 responses per minute).

Overall, the high levels of calibration accuracy and precision indicate that either the equipment used for recording was good for the current purpose, or that the observational task was easy, or a combination of both. DataPal does replicate the ideal form of electronic recording forecast by Johnston and Pennypacker (1980). Experienced observers and some novices held their hands over the laptop keyboard so they were able to type the codes as they detected target responses without having to look away from the video monitor showing the criterion (i.e., calibration or reference) samples. Further calibration study will be required before it can be determined whether varying the observer's recording response affects accuracy (e.g., with touch-screen recording devices).

Regarding ease of the observers' task, as mentioned earlier, the scripts used for creating the calibration samples were derived from real archived clinical and research data collected by experienced observers using DataPal. In comparison to the recording protocol used in our study, which had three active event keys, the archival observational protocols on which our video samples were based typically had between seven (five events, two durations) and 16 (nine events, six durations) active keys. The present observational protocol can be seen as less demanding than the experienced observers' usual task. The effects of increased complexity of recording demands may be investigated with future calibration studies.

We presented the calibration statistics (Table 2) and also the percentage agreement between observers' records and criterion records using common interobserver agreement algorithms (Table 1). Both types of data (calibration statistics and percentage agreement with the criterion records) may be useful for evaluating the accuracy of direct observation. The agreement percentages show that observers' code-recording responses were largely under stimulus control of target responses that occurred on the monitor. It is conceivable for accurate and precise summary measures of rate by observers' recording to be controlled by irrelevant stimuli (e.g., the observer recorded 10 responses, the same as the reference value, but the observer was measuring some other nontarget response that occurred at different times from the target response; Baer, 1977). That could not be detected from calibration analysis with whole-session summary data. Hence, inclusion of agreement measures in calibration studies with direct behavioral observation should continue as an additional indicator of the quality of data.

In addition to showing that calibration methods used in other natural sciences can be used to assess the accuracy and precision of direct observation measures, it also is important to discuss the potential uses and benefits of calibration for applied behavior analysis. First, as shown here, the quality of individual observers as components of a measurement system can be assessed more thoroughly using calibration with known values than by using traditional interobserver agreement methods. Assurances regarding observer's accuracy, or indications for further training, arise directly from calibration. Second, variations in measurement systems can be compared using scientifically sound calibration methods. One such comparison that should be the focus of future investigation is the accuracy and precision of direct observation data collected using laptop computers versus pencil-and-paper methods. Third, calibration methods can be developed that would allow researchers to report routinely on the accuracy and precision of the data they present in research publications. To do so would probably require the development and validation of a less labor-intensive method than used in the present study. Fourth,

improvement in data-based decisions can be foreseen. For example, if the present data were from a clinical research study on reducing pinching, one could be assured that observers' measures that were different by more than twice the confidence intervals for predictions were more likely due to real differences in the levels of pinching than to random observer error (i.e., lack of precision). Fifth, scientists from other branches of natural science with whom behavior analysts collaborate should recognize the employment of calibration as being in line with methods for assessment and improvement in measurement in their fields.

The potential benefits of calibration for applied behavioral researchers are hypothetical. Any such speculations need to be thoroughly evaluated through research that demonstrates to our field the benefits and disadvantages of a variety of methods for obtaining the reference values that are essential in calibration. An important question concerns the circumstances in which it might be reasonable to forecast that calibration could replace interobserver agreement as the accepted method for quantifying the quality of our data. Consider an applied study in which the baseline rate of responding is in the range of 6 to 8 responses per minute. Reference values for two or more sessions during baseline could be obtained from criterion records. Reference values would then be obtained from several additional sessions while the rate of responding was decreasing during treatment (e.g., between 6 and 1 responses per minute) and again when low levels of responding occurred in maintenance phases (e.g., less than 1 response per minute). Up to 10 sessions' reference values would be required to compare with observers' measurements for calibrating the study's data. The most economical method for obtaining reference values would be to employ a single expert observer to measure reference values for calibration (as in Sanson-Fisher et al., 1980; Wolfe et al., 1986). The observers' data would be presented graphically as the substantive data from the study (as it is presently done) along with calibration statistics, either numerically or graphically, instead of the familiar interobserver agreement measures. Whether behavior-analytic researchers, reviewers, and editors would accept such a method is an empirical question.

An important question regarding the future of calibration in applied behavior analysis is: How ''true'' are the reference values derived from the criterion records for the purpose of calibration? The impossibility of determining the accuracy of observations has been argued by Gresham (2003):

> Unlike measurement in the physical and biological sciences … there is no gold standard against which to compare an observer's recording of behavior and environmental events to the ''true'' state of nature … there seems to be no entirely defensible way of establishing the accuracy of [measurements in] FBA [functional behavioral assessment]. (p. 288)

Two measures are required to determine responses per minute: count and time. Because the calibration samples used in the current study were (and are) digitized video, in principle, any skeptical reader could view them and count the number of pinches as defined. Count is an absolute measure, with no systematic or random error. Skeptics could, in principle, test the accuracy and precision of the timing of 10 min (600 s) in the video editing software employed against the standard atomic clock. We doubt that calibration of timing would alter our findings on observer accuracy and precision ''in relation to the requirements of their use'' (Hibbert & Gooding, 2005, p. 2). Further, data derived from criterion records created in the manner described in this investigation may well be ''the accepted reference value'' (The Royal Society of Chemistry, 2003, p. 1) obtained from ''an exemplar method'' (Doebelin, 1966) in the field of applied behavior analysis.

The present study calibrated observers over a relatively wide range of rates of responding. Because the mean response rate for free-operant

responding in behavioral research has been 0.9 responses per minute (Mudford et al., 2011), exploratory calibration studies that focus on a narrower range (e.g., 0 to 2 responses per minute) should be conducted. When the range of rate of responding is small across a set of calibration samples, accuracy and precision can be investigated in finer detail because, with precise observers, confidence bands for predictions narrow further than could be illustrated here. Conversely, calibration studies at higher rates of responding could be conducted to include 90% of the maximum rates reported in applied behavior-analytic research, (i.e., up to 24 responses per minute, Mudford et al., 2011). Other studies may demonstrate methods for calibrating measures of (a) duration of responding, (b) interresponse times and latencies, and (c) sequences of different events to assess stimulus–response or response–response relations. These are just some examples.

There are some caveats regarding further studies in calibration in applied behavior analysis that might follow from our example. First, simple linear regression may not be the appropriate analysis model for all calibration studies in applied behavior analysis. For example, although the obtained relation between criterion reference values and observers' measures may be nonlinear, the observers' data may still be judged to be adequate for the purpose. Second, we have shown how predictions of Y values from X values can be used to evaluate whether a particular measurement system might be suitable for its purpose. Predicting X (true values) from Y (obtained values) requires additional analysis known as *inverse regression* (The Royal Society of Chemistry, 2006), which is less often available in commonly used data-analysis software. Third, we have addressed only a subset of potentially relevant metrological concepts used in other natural sciences.

Our study was designed specifically as an initial demonstration of calibration for quantifying the quality (specifically, accuracy and precision) of direct observational data. The study's contribution to the behavior-analytic literature may be to introduce a new line of research akin to that concerning interobserver agreement that commenced in the early 1970s. We anticipate further research that elucidates the advantages and disadvantages of calibration. Further studies could assess the use of calibration for methodological purposes (e.g., comparing different behavior-recording methods), to assess training of observers to record accurately, and for practical purposes (e.g., to show the accuracy and precision of data from intervention studies).

REFERENCES

Baer, D. M. (1977). Reviewer's comment: Just because it's reliable doesn't mean that you can use it. *Journal of Applied Behavior Analysis*, *10*, 117–119.

Boykin, R. A., & Nelson, R. O. (1981). The effect of instructions and calculation procedures on observers' accuracy, agreement, and calculation correctness. *Journal of Applied Behavior Analysis*, *14*, 479–489.

Bragg, G. M. (1974). *Principles of experimentation and measurement*. Englewood Cliffs, NJ: Prentice Hall.

Burke, S. (2001). *Regression and calibration*. LC•GC Europe Online Supplement. Chromatography Online. Retrieved from http://chromatographyonline.findanalytichem.com/lcgc/data/articlestandard/lcgceurope/502001/4500/article.pdf

Cooper, J. O., Heron, T. E., & Heward, W. L. (1987). *Applied behavior analysis*. Columbus, OH: Merrill.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Pearson.

Doebelin, E. O. (1966). *Measurement systems: Application and design*. New York: McGraw-Hill.

Fahmie, T. A., & Hanley, G. P. (2008). Progressing toward data intimacy: A review of within-session data analysis. *Journal of Applied Behavior Analysis*, *41*, 319–331.

Feigelson, E. D., & Babu, G. J. (1992). Linear regression in astronomy. II. *The Astrophysical Journal*, *397*, 55–67.

Garofolo, F. (2004). Bioanalytical method validation. In C. C. Chan, H. Lam, Y. C. Lee, & X. Zhang (Eds.), *Analytical method validation and instrument performance verification* (pp. 105–138). Hoboken, NJ: Wiley.

Gresham, F. M. (2003). Establishing the technical adequacy of functional behavioral assessment: Conceptual and measurement challenges. *Behavioral Disorders*, *28*, 282–298.

Grimm, H., Gilfanov, M., & Sunyaez, R. (2003). High-mass X-ray binaries as a star formation rate indicator in distant galaxies. *Monthly Notices of the Royal Astronomical Society*, *339*, 793–809.

Hauck, W. W., Koch, W., Abernethy, D., & Williams, R. L. (2008). Making sense of trueness, precision, accuracy, and uncertainty. *Pharmacopeial Forum*, *34*, 838–842.

Hibbert, D. B., & Gooding, J. J. (2005). *Data analysis for chemistry: An introductory guide for students and laboratory scientists*. New York: OUP.

Holland, J., Dabelsteen, T., Bjorn, C. P., & Pedersen, S. B. (2001). The location of ranging cues in wren song: Evidence from calibrated interactive playback experiments. *Behaviour*, *138*, 189–206.

Hong, L. E., Avila, M. T., Wonodi, I., McMahon, R. P., & Thaker, G. K. (2005). Reliability of a portable head-mounted eye tracking instrument for schizophrenia research. *Behavior Research Methods*, *37*, 133–138.

Johnston, J. M., & Pennypacker, H. S. (1980). *Strategies and tactics of human behavioral research*. Hillsdale, NJ: Erlbaum.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hillsdale, NJ: Erlbaum.

Johnston, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of behavioral research* (3rd ed.). New York: Routledge.

Kapust, J. A., & Nelson, R. O. (1984). Effects of the rate and spatial separation of target behaviors on observer accuracy and interobserver agreement. *Behavioral Assessment*, *6*, 253–262.

Kramer, R. (1998). *Chemometric techniques for quantitative analysis*. New York: Dekker.

Lee, N. M., & Hockey, P. A. R. (2001). Biases in the field estimation of shorebird prey sizes. *Journal of Field Ornithology*, *72*, 49–61.

Lerman, D. C., Tetreault, A., Hovanetz, A., Bellaci, E., Miller, J., Karp, H., et al. (2010). Applying signal-detection theory to the study of observer accuracy and bias in behavioral assessment. *Journal of Applied Behavior Analysis*, *43*, 195–213.

Mudford, O. C., Locke, J. M., & Jeffrey, K. (2011). Rates of responding measured by continuous recording in applied behavioral research. *Behavioral Interventions*, *26*, 41–49.

Mudford, O. C., Martin, N. T., Hui, J. K. Y., & Taylor, S. A. (2009). Assessing observer accuracy in continuous recording of rate and duration: Three algorithms compared. *Journal of Applied Behavior Analysis*, *42*, 527–539.

Mudford, O. C., Taylor, S. A., & Martin, N. T. (2009). Continuous recording and interobserver agreement algorithms reported in the *Journal of Applied Behavior Analysis* (1995–2005). *Journal of Applied Behavior Analysis*, *42*, 165–169.

Niemoller, A., & Behmer, D. (2008). Use of near infrared spectroscopy in the food industry. In J. Irudayaraj & C. Rey (Eds.), *Nondestructive testing of food quality* (pp. 67–118). Ames, IA: Blackwell.

Powell, J., Martindale, B., Kulp, S., Martindale, A., & Bauman, R. (1977). Taking a closer look: Time sampling and measurement error. *Journal of Applied Behavior Analysis*, *10*, 325–332.

The Royal Society of Chemistry. Analytical Methods Committee. (2003). *AMC technical brief: Terminology—the key to understanding analytical science. Part 1: Accuracy, precision and uncertainty. (AMCTB No. 13)*. Retrieved from http://www.rsc.org/images/brief13_tcm18-25955.pdf

The Royal Society of Chemistry. Analytical Methods Committee. (2006). *AMC technical brief: Uncertainties in concentrations estimated from calibration experiments (AMCTB No. 22)*. Retrieved from http://www.rsc.org/images/Brief22_tcm18-51117.pdf

Russell, B. (1937). *The principles of mathematics* (2nd ed.). New York: Norton.

Sanson-Fisher, R. W., Poole, A. D., & Dunn, J. (1980). An empirical method for determining an appropriate interval length for recording behavior. *Journal of Applied Behavior Analysis*, *13*, 493–500.

Wolfe, V. V., Cone, J. D., & Wolfe, D. A. (1986). Social and solipsistic observer training: Effects on agreement with a criterion. *Journal of Psychopathology and Behavioral Assessment*, *8*, 211–226.

Yang, W., Gu, D., Chen, J., Jaquish, C. E., Rao, D. C., Wu, X., et al. (2008). Agreement of blood pressure measurements between random-zero and standard mercury sphygmomanometers. *American Journal of the Medical Sciences*, *336*, 373–378.

Received June 24, 2010
Final acceptance November 9, 2010
Action Editor, James Carr